Predictions from Physical Fitness Tests
Impact of Age and Gender

U. Bergh and U. Danielsson
Defence Research Establishment, SE-172 90 Stockholm, Sweden.
Phone: +46 8 706 3210, Fax: +46 8 706 3309, email: ubergh@sto.foa.se

Summary
Physical fitness tests are employed in most armed forces; the purpose being to avoid persons with insufficient fitness. The predictive value is strongly influenced by the prevalence of the tested quality. In regard to physical work capacity, higher values are more prevalent among males compared to females and among younger people compared to older ones. At a prevalence of .9 for males and .4 for females, the success rate among those who passes the test would theoretically be 95% and 70%, respectively. Prevalence should be included when predicting the possible outcome of different tests. This theoretical example is in line with empirical findings. For example, among fire-fighters who had passed a treadmill test, the success rate in a smoke-diving task was 90% in age group 20-30 years, 78% in age group 31-40 years, 69% in age group 41-50 years, and 30% in age group 51-60years.

Introduction
Most jobs in the armed forces are including physical work. Therefore, tests of physical fitness are employed; the purpose being to avoid persons with insufficient fitness. Different principals are used when setting the pass/fail level. One is to have the same requirement irrespective of age and gender, another is to adjust the demands to those factors.

Ideally, the test should make all those having a sufficient capacity pass, while the others should fail. Such test are, however, practically non-existent, meaning that some with an insufficient capacity will pass, while some of those having a sufficient capacity will fail. This lack of perfection will induce errors, some of which may become quite costly.

An often forgotten problem in this context is that the task success rate among those passing a test varies with the true capacity of the population from which the group is recruited. A high prevalence of people having a sufficient capacity will give a higher success rate, among those who pass the test, than if the group is coming from a population with a lower capacity.

A high level of physical performance is more prevalent among men compared to women and among young adults compared to middle-aged ones.

Test reliability is considered very important. Some of its implications are, however, not frequently recognized.

The purpose of this paper is to elucidate some fundamental, but often forgotten theoretical principals and their impact on the selection process especially in regard to age and gender.

Test scores and task performance
Relating test scores to task, or job, performance very often display far from perfect relationships. This is illustrated in figure 1. Including minimum requirements for the job and pass level for the test will divide the population into four categories:

- True positive; passing the test - sufficient capacity
- True negative; failing the test - insufficient capacity
- False positive; passing the test - insufficient capacity
- False negative; failing the test - sufficient capacity

The first two outcomes are correct, while the latter two are erroneous. These errors differ from one another regarding the consequences they produce. The false positive results will lead to an acceptance of people with insufficient capacity, who eventually may leave the organization and that represents a cost giving little or no benefit. The false negative results give rise to rejections of qualified people. In turn it may lead to problems in finding enough people with sufficient quality.

Increasing the pass level will reduce the problem associated with recruiting people with insufficient capacity, but it will increase the problem with

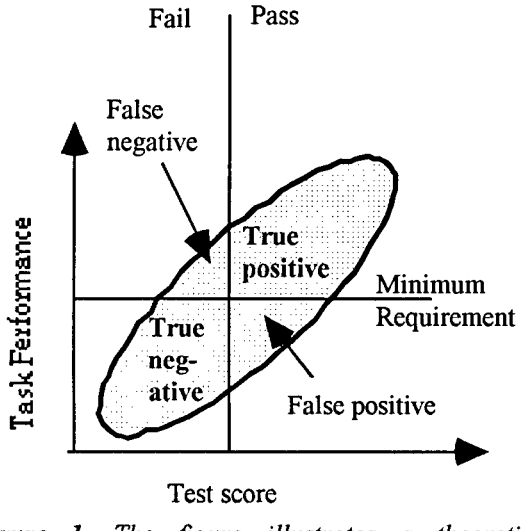

Figure 1. The figure illustrates a theoretical relationship between test scores and task performance.

Paper presented at the RTO HFM Workshop on "Officer Selection", held in Monterey, USA, 9-11 November 1999, and published in RTO MP-55.

rejecting people with sufficient qualities (see figure 1). So, with a given test, it is not possible to reduce both of these errors. It is a question of analyzing the effects of the errors as to minimize their negative consequences.

So far, the qualitative aspects have been discussed. In order to better understand the practical importance, it is necessary to look at some of the quantitative aspects.

Prediction, Sensitivity, Specificity and Prevalence
First, we should define those terms. The **positive predictive value** is the fraction of truly qualified among those who passes the test.
Sensitivity is the probability that the test will identify a given quality. If the sensitivity is unity, the test will find all with that quality; a value of .5 means that only half of the people having that quality will be identified by the test. **Specificity** is the probability that the test will identify people who are lacking this quality. A specificity of 1.0 means that the test is expected correctly identify all of those who are lacking that quality, and that nobody who is lacking the quality would be judged as having it. **Prevalence** is the fraction of a population having a certain quality.

The numeric relationship between these components is:

$$PPV = Pr \cdot Sens/(Pr \cdot Sens + (1-Pr)\cdot(1-Spec)) \quad \text{(eq. 1)}$$

where
PPV = Positive predictive value,
Pr = Prevalence,
Sens = sensitivity,
and Spec = Specificity.
So, increasing Pr, Sens or Spec work in the direction of increasing PPV. In other words, given the values for sensitivity and specificity, the fraction of truly qualified among those passing the test increases if this certain quality becomes more prevalent in the population. Hence, one would expect that among those passing a strength test, more men than women, and more 30 year old than 60 year old ones, would be able manage jobs that include the lifting of heavy objects. The order of magnitude of this effect is primarily influenced by the specificity of the test. This is because decreasing specificity leads to an increasing denominator of equation 1, while changing the sensitivity affects both the numerator and denominator.
Knowing the numeric value of these factors, it is possible to calculate and predict the outcome of various test procedures. An example is given in figure 2, which is displaying the outcome of test performed on different populations which differs in regard to the prevalence. Assuming Sens = Spec = .8, the PPV is .63 for Pr = 30%, and .94 for Pr = 80%.

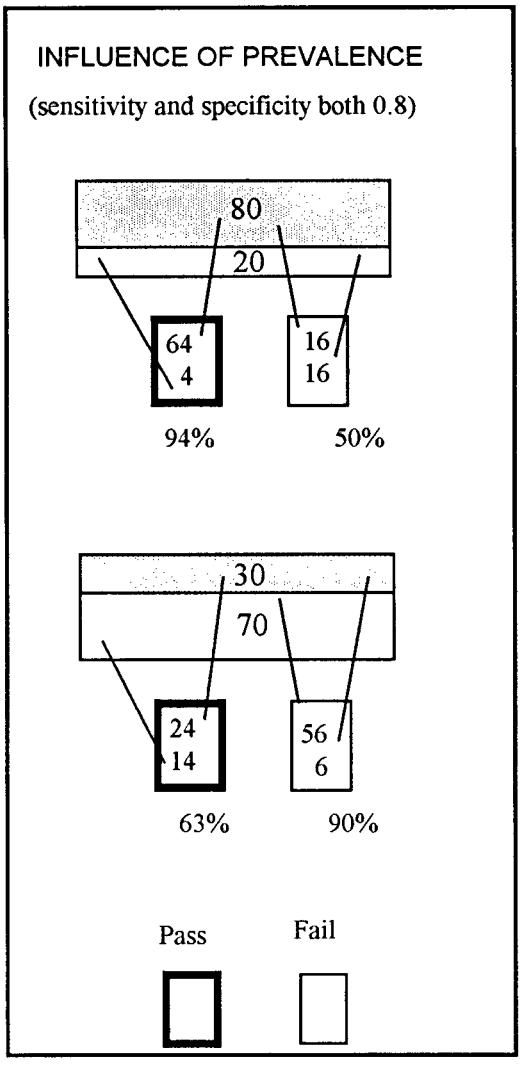

Figure 2. Influence of prevalence on the predictive value of a given test. Numbers expressed as per cent denote the fraction of correct predictions among those who pass as well as among those who fail.

Another example is given in figure 3, displaying the effect of prevalence for tests with different sensitivity and specificity. It is evident that the effect of prevalence is greater for poor test than for good ones.

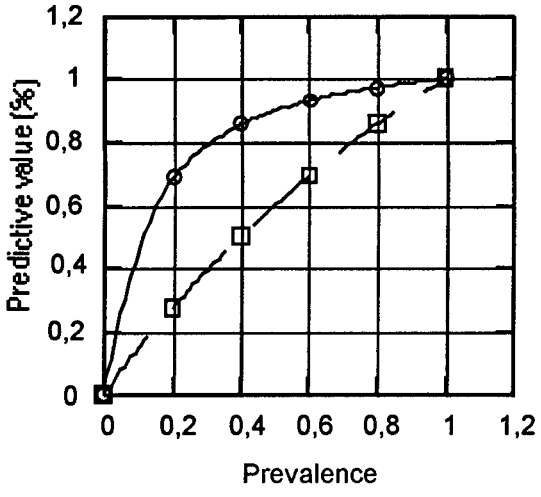

Figure 3. *This figure demonstrates how the predictive value is influenced by the prevalence and the quality of the test. It is assumed that the better of these tests is characterised by sensitivity = specificity = .9 while the poor test had sensitivity = specificity =.6*

These effects can be quite dramatic as illustrated in figure 4, Among men the PPV was 95% and among women 70%, i.e., among those who passed the test 95% of the men is expected to have a sufficient capacity compared to 70% of the women.

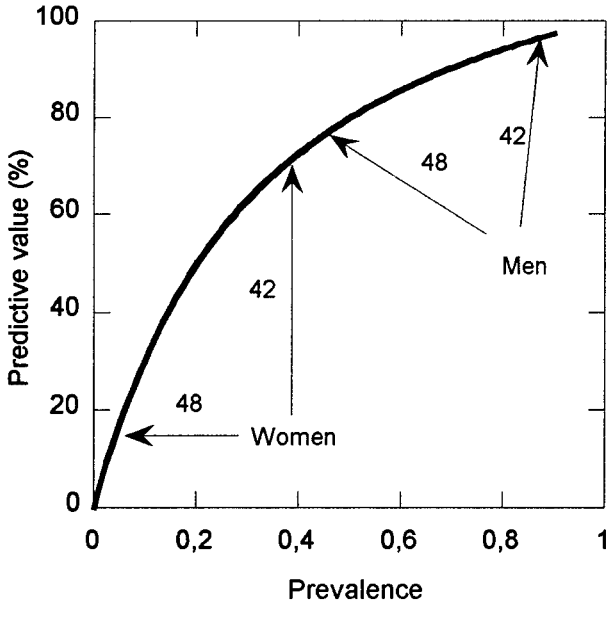

Figure 4. *This figure demonstrates that prevalence strongly influences the predictive value. Also, the prevalence of different levels (42 and 48 ml•kg^{-1}•min^{-1}) of maximal oxygen uptake among men and women of age 20 to 25 years is indicated.*

The assumption that if you pass the test your physical fitness is adequate is not always correct, and moreover the magnitude this lack of correctness varies with the prevalence.

Another consequence of this is that in order to correctly identify people with characteristics that are less prevalent, one needs test with very high sensitivity and even more important a very high specificity.

The PPV describes only one type of error; the order of magnitude of which is affected primarily by the specificity of the test and the prevalence.

The sensitivity is of course important because tests with low sensitivity will fail to identify a lot of people with sufficient capacity (see figure 1).

Age

Physical performance decreases with age among adults. Hence, the number of people able to manage a given physical task is lower among the 40 year old than in the 25 year old. The same effect is, thus, expected among different age groups.

Another interesting question is the order of magnitude, i.e., what fraction of different age groups are able to achieve given levels of performance. An example, is presented in figure . The calculation is based on the results from 371 officers running 2000 m and a reduction in maximal oxygen uptake by 8% per 10 years (Shwarts & Reibold, 1990).

The effect is quit dramatic; the fraction increased from 0 % at age 25 to 40% at age 40. Thus, almost half of the population will display an inadequate capacity.

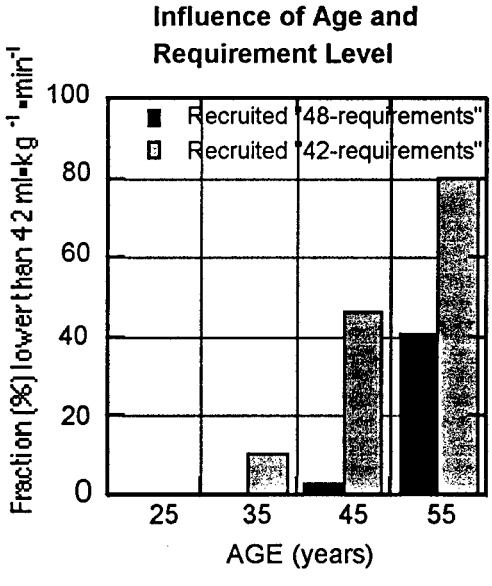

Figure 5. *Influence of age and recruitment requirement level on the fraction that will not be able to pass a given test.*

This may cause unwanted consequence for the individual as well as for the organization. One way to

reduce these problems is to increase the requirements at the age of 25, leaving enough room for age-related reductions of physical performance. The side effect is, however, that more people will be excluded, some of which having qualities of importance for the job. Moreover, it will become much more difficult to recruit women. For example, increasing the entrance requirements to a level that would reduce the fraction of men that fails to 5 % at age 55, would exclude 95-98% of the women.

Success rate and test scores.

In general, the correlation between test score and actual job performance is fairly low. Thus, test score variance usually explain less than 25 percent of the variation in performance, which is not too impressive. One might even ask if testing is useful.

This leads to the need of defining the term useful. First, the tests have to produce better predictions than using pure chance, e.g., a lottery. Second, the cost for testing must be less than the costs of the problems resulting from an inadequate selection.

However, providing that the test produces a result better than pure chance, test with moderate quality may be useful. For example, the correlation between running performance and field exercises is seldom very high (see figure 8) Still the fraction failing to manage such exercises is usually lower among subjects with high running performance than among those showing low values (figure 9). However, the cost of such testing must still be weighed against its benefit.
Information about different success rates can be treated statistically (Lubinski & Humphreys, 1996).

Reliability and classification

Another aspect is test reliability, i.e., capability of a test to reproduce the results. For example, measuring a person's body mass twice should give very much the same result. During selection procedures, the test results are often used for the purpose of dividing the population into different groups, e.g. "low, "average", "high". Ideally, two consecutive tests should place the subject in the same group at both trials. In reality, such an outcome is very rare. Most tests perform less well. There are mainly two factors that influence the performance of a test in that respect (figure 6)

- number of groups; more groups will result in less subjects getting the same classification at both trials.
- reliability coefficient; higher value gives more subjects getting the same classification at first and second trials.

A practical implication is that the number of groups has to be adjusted to the actual level of test reliability; otherwise the classification will result too much from chance and too little from true value.

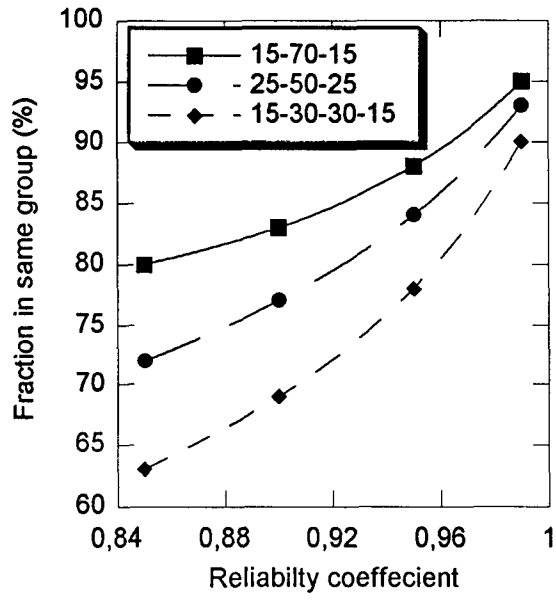

Figure 6. This figure illustrates the effect of test reliability on fraction of the tested population that will get the same classification when test twice. The influence of different classifications are also shown. The figures 15-70-15 denote the fraction of the population in each group. (Data adopted from A. Avén, 1996)

Practical results and implications

This can be illustrated by a study on firefighters of different age, who performed a walking test on a treadmill (Danielsson & Bergh, 1997). Those who passed also performed a smoke-diving test. The success rate in that test was higher among the younger age groups compared with the older ones (figure 7). This is an illustration of the principals described earlier, i.e., that among people that have passed a test, the success rate in job-related tasks is lower in populations with a lower capacity (lower prevalence of people with high capacity). A practical significance of this is that tests may produce erroneous information about the chances of managing the job.

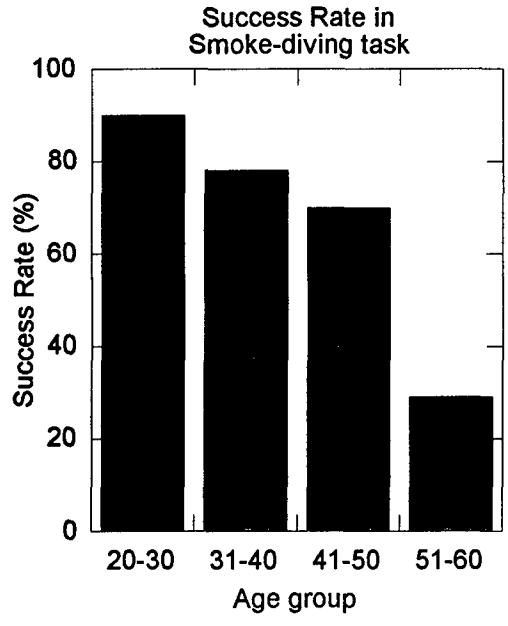

Figure 7. Success Rate in a smoke-diving task among firefighters that have passed the compulsory treadmill test.(Data from Danielsson & Bergh 1997)

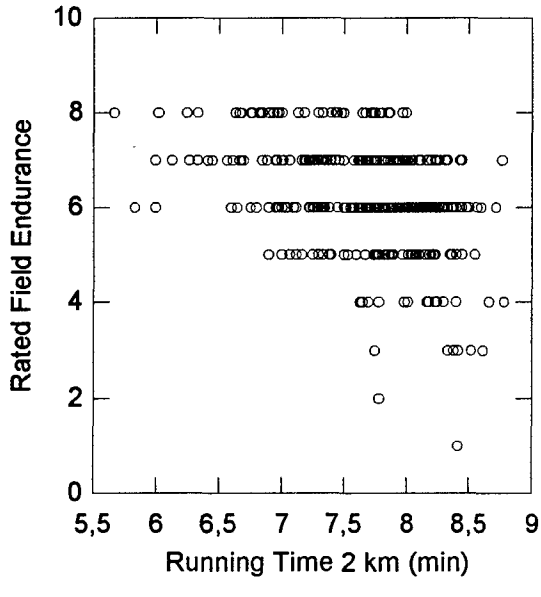

Figure 8. Rated Field Endurance in relation to running time for 2 km. Adequate endurance corresponds to a rating score of 5.

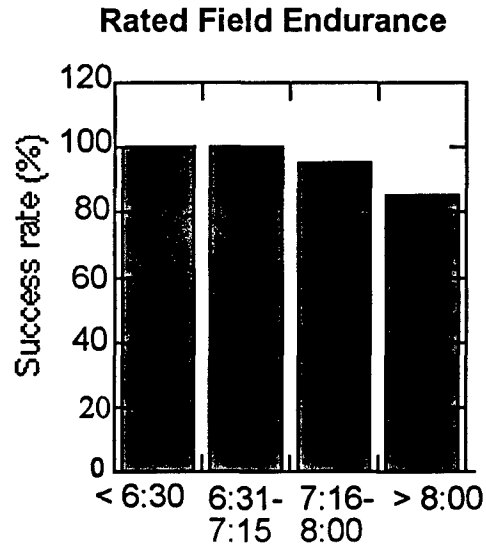

Figure 9.The fraction rated as having adequate field endurance among groups with different running performance. Note that even if the correlation shown in the previous figure is rather low there is a difference in rated field endurance between groups.

References

Avén A. Isokaiprovets reliabilitet (The reliability of the Isokai test) PM 47:54, 1996.

Danielsson U. and Bergh U. Fysiska krav på befattningar inom räddningstjänsten (Physical work demands in the Rescue Service). FOA-R-97--00549-720--SE, ISSN 1104-9154, 1997.

Lubinski D. and Humphreys L. G. Seeing the forest from the trees. Psychology, Public Policy, and Law Enforcement, vol. 2, No 2, 363-376, 1996

Shvartz E. and R. C. Reibold. Aerobic fitness norms for males and females aged 6-75 years: a review. Aviat. Space Environ. Med. 61: 3-11, 1990.